GEDII GENDER DIVERSITY IMPACT –
Improving research and innovation
through gender diversity

WWW.GEDII.EU

Addendum to Project Deliverable D4.3

Documentation of Modifications/Corrections between D4.3 version 1.0 published 23rd March 2018 and D4.3 version 2.0 published 15th January 2019

13th of September 2019

This report has been compiled by Jörg Müller

Its contents has been approved by:

Jörg Müller
Anne Laure Humbert
Elisabeth Guenther

Project Consortium

Universitat Oberta de Catalunya, Spain
Project Coordination
Dr. Jörg Müller
Dr. Milagros Sáinz
Dr. Rosa Borge
Dr. Julià Minguillón
Dr. Julio Meneses
Dr. Sergi Fàbregues

Hochschule Furtwangen, Germany
Prof. Dr. Ulrike Busolt
Sandra Klatt
Wiebke Kronsbein

Örebro University, Sweden
Dr. Ulf Sandström
Dr. Anne-Charlott Callerstig
Prof. Dr. Liisa Husu

Verband der Elektrotechnik Elektronik Informationstechnik e.V., Germany
Dr. Walter Börmann
Nicole Bienge
Andreas Landwehr
Annika Gereke

Oxford Brooks, UK
Dr. Anne Laure Humbert
Dr. Elisabeth Guenther

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 665851.

Introduction

This document is a complementary file to Deliverable 4.3 of the GEDII project. It summarizes the changes and corrections between version 1 and version 2 of "D4.3 Survey Analysis and Performance Indicator Research Report". A first version of the report was produced on 24th of March 2018 and send to the European Commission. This first version of the report is available as

- D4.3_GenderPerformance_23MAR2018.pdf

After the completion of the first version several issues with the analysis were discovered that required correction. A second, revised version of the report was published on the 15th of January 2019. This second version of the report is available as

- D4.3_GenderPerformance_15JAN2019.pdf

The results published in D4.3 are complex in terms of the data used, data pre-processing steps involved as well as the actual analysis carried out. Modifications have been introduced at different points in the data-processing pipeline. A short summary of the overall architecture and logic of analysis will make it easier to understand the track changes described further down.

The analysis combines data from two different sources: one the one hand, GEDII carried out a cross-country survey of R&D teams across Europe. The survey file contains important socio-demographic information about team members as well as other team related variables such as team climate, power relations, leadership style among others. In order to achieve a high response rate, four different delivery mechanisms of the same questionnaire were offered; much of the pre-processing scripts combine these four single survey files into one. On the other hand, bibliometric performance data was compiled for each team from Web of Science and the Patstat database. Compilation of performance data for each R&D team is itself a complicated process, involving author disambiguation or definition of adequate time windows for inclusion of articles. In the final analysis, the survey responses and the performance data has been merged. Changes have been introduced on almost all steps of the data-processing pipeline between D4.3 v1 and D4.3 v2, most importantly:

- cleaning (involving aggregation and re-codification) of variables in the cross-country survey
- software packages used in running the analysis on the link between gender diversity and bibliometric performance
- transformations and set of variables used in the analysis (both dependent and independent variables)
- number of groups included in the analysis
- treatment of missing cases
- transformation and interpretation of results
- definition (time-window) of dependent variables

In what follows, a more detailed summary of the introduced changes will be presented. Whereas provided R scripts (build_ccs.R and build_tld.R, see below) clearly document how “raw” data files downloaded from the online survey platform are transformed into the data matrix used for statistical modeling, there is no such documentation available for the processing of bibliometric performance data, i.e. how bibliometric information downloaded from the Web of Science for each author has been cleaned and processed to yield the group level performance indicators “FAP” and “PMM”.

Overview of available files

All referenced files in this document are provided in an accompanying compressed archive “ChangesD4.3JAN2018-15JAN2019.zip”. This includes the processing scripts for cleaning the “raw” survey data as well as the scripts used for the analysis. We also provide two Docker files for creating the software environment for version 1 and version 2 of this reports respectively.

Concerning the the survey data: a public version of the survey result data package is available:

> Müller, J., Busolt, U., Callerstig, A.-C., Guenther, E. A., Humbert, A. L., Klatt, S., & Sandström, U. (2019). *GEDII Survey on Research Teams Dataset* [Data set]. https://doi.org/10.5281/zenodo.2545196

Please note, however, that the individual level responses have been removed out of data protection concerns.

Reproducible Example of analysis used in D4.3 version 1

To start with, a reproducible example of the analysis carried out in D4.3 version 1 has been provided (see “Ranalysis-JAN2018-rocker/Dockerfile”). The delivery of such a reproducible example is not straight forward since the used software packages (R Project for Statistical Computing)[1] are constantly updated. Although the scripts and data used to produce the analysis is fixed, the software environment used in January 2018 had been superseded in January 2019; re-running the analysis with the same code and data produces different results since underlying software packages had been error corrected and updated in the meantime[2].

In order to reproduce the software environment available for R in January 2018 when the first analysis for D4.3 was carried out, UOC has used an approach involving “Docker”[3]. Docker and the R-specific repository “Rocker” basically allows users to install previous versions of R packages in a stand-alone

[1] https://www.r-project.org/

[2] These issues and other are known in the Open Science Community which also provides examples for addressing these. See for example the framework and tools at https://github.com/benmarwick/rrtools including references.

[3] For a short introduction to the overall approach see https://colinfay.me/docker-r-reproducibility/ For documentation on Docker see https://docs.docker.com/ for installing specific r-base environments on Docker see https://www.rocker-project.org/ for installing “outdated” R packages see https://mran.microsoft.com/timemachine

"container" and re-run a certain analysis with "outdated" software packages. This has been done for the R environment 28 of January 2018 (see "Ranalysis-JAN2018-rocker/Dockerfile") which generates the file "Ranalysis-JAN2018-rocker/d43v1.html". Providing such a reproducible example is important because one of the errors detected in version 1 of the report concerns the formatting of result tables with the function "sjt.lm()" of the sjPlot package. Since the "sjt.lm()" function is meant to format linear regression models, it inserts under its default settings a R² value in the formatted table output. However, since our analysis uses negative binominal model, the R² values are not related to the underlying model and incorrect.

Comparing the output generated by the R scripts used in January 2018 ("d43v1.html") with the data published in D4.3 version 1 shows that the result match.[4] Most importantly, the erroneous R² coefficients match for all models. As a consequence we can confirm that no results have been willfully manipulated or results faked. Allegations of manipulation of results used in D4.3 version 1 are therefore unsubstantiated.

The results of the modeling, including the erroneous R² values for the models were available to the Consortium as a draft chapter for D4.3 since January 2018 in preparation of the 5th Project Meeting in Barcelona (29th of January 2018). The error was not detected, neither by the authors of the chapter nor by any other Consortium member until September 2018, when it was pointed out by Ulf Sandström.

The resulting revision of the analysis and scripts used for D4.3 version 1 correctly identified the problematic use of the "sjt.lm()" function for the formatting of result tables among other issues described in this document (and summarized and documented in an email exchange among Consortium members in October 2018). The errors were corrected in order to produce version 2 of D4.3

The "sjt.lm()" function has been deprecated in newer versions of the "sjPlot" package and has been replaced by "sjPlot::tab_model()" function which can be used for formatting result tables. In contrast to its predecessor, the "tab_model()" detects a wide variety of different underlying models (linear model or negative binominal among others) and adjusts its display accordingly. No warning message is displayed by "sjt.lm()" function when used incorrectly with other than linear regression models.

Summary of changes/corrections for D4.3 version 2

During the process of revising and correcting D4.3 several other modifications were introduced either to the data itself, the pre-processing of data, the aggregation levels of variables, the actual analysis or interpretation of results. During the production of D4.3 version 2, UOC revised the names of all variables and provided extensive documentation of all variables as R data package "gediiccs.private". **A public version of the same software/data package is available (see above); however, the individual level responses have been removed out of data protection concerns.** The public version

[4] Note that the order of the result tables in the HTML file does not follow the order of the models and tables in the D4.3 report.

contains the team level aggregated data only, apart from the complete pre-processing and variable documentation.

Below, we present a summary of the main changes. For a complete documentation of the pre-processing steps carried out in the generation of the survey data file, inclusion of performance data and aggregation of data on the team level see

- survey-preprocess-scripts/build_gediiccs.R
- survey-preprocess-scripts/build_tld.R

These files generate the data matrix in the “gediiccs.private” (or “gediccs” public) package, used for the performance analysis and reporting of results. Detailed summary of changes introduced are available in the

- survey-preprocess-scripts/CHANGELOG

file.

Data pre-processing and aggregation

Main changes introduced to the pre-processing of survey data. See also documentation of variables in “gediiccs.private” package.

- Control of missing cases (excluded from analysis) is done manually, whereas it was done automatically by the “glm.nb()” function previously. Version 2 has N=88 compared to version 1 with N=81. Based upon performance data but also other decisions, some groups were excluded, others incorporated.
- Gender data has been complemented in such a way as to have gender assigned to all team members (not just those responding to the survey) based upon the first name of each team member.
- Gender of leader has been manually assigned (first name, manual web lookup)
- The calculation of the “Power Influence Disparity” has changed
- Calculations of team size were tested based upon two different approaches to identify seniors in the group: based upon a) questionnaire data or based upon b) bibliometric data. This gives rise to different classification of “Team types”, according to the number of senior researchers in the group.
- The Gender Diversity Index has been re-calculated based upon the modifications introduced in the groups.

Performance data

Bibliometric performance data was updated twice, 12th and 28th of November 2018, including the following operations:

- Performance data for some groups was manually corrected (groups: 30, 39, 47, 104, 56, 122, 123) including a correction for lowest score (previous “0” now “0.01”)
- Exposure time for bibliometric records has been reduced from 8 to 5 years.

In an earlier update, performance indicators (FAP and PMM) already were multiplied by 1000 in order to avoid rounding errors of dependent variables (negative binominal expects count data, i.e. whole integers).

Analysis & Interpretation

- Result coefficients of models are interpreted as “Incidence Rate Ratios”
- R package “sjPlot” with function “sjPlot::sjt.lm()” has been updated. The “sjt.lm()” function has been deprecated and is not available anymore. Correct function for formatting negative binominal “glm.nb()” is “sjPlot::tab_model()” which detects the underlying negative binominal model and formats it accordingly.

Reproducible Example of analysis used in D4.3 version 2

Following the same logic for providing a reproducible example of the analysis carried out for D4.3 version 1, a Dockerfile for January 15th of 2019 is provided (see folder Ranalysis-JAN2019-rocker) . Running the Docker image will produce the HTML result tables (Chapter3_AnnexII_13JAN2019.html) used in D4.3. version 2.